**Designing a Credit Risk Policy for Startup Lending**

**Client:** Major Global Bank

**Consultants:** Anastasia De Bondt, Arthur, Kasper, Hannah Staudenherz

**Date:** May 2025

**1. Introduction**

This report presents a data-driven strategy to design a more robust credit risk policy for evaluating startup loan applications. Using historical data from previously funded startups, we identify key patterns and risk factors associated with startup failures. The proposed methodology leverages segmentation, funding timelines, and sectoral analysis to improve decision-making, reduce default rates, and increase the return on investment for the client.

**2. Data Overview**

The dataset includes key attributes:

* **Identification:** permalink, name, homepage\_url
* **Financial:** funding\_total\_usd, funding\_rounds
* **Temporal:** founded\_at, first\_funding\_at, last\_funding\_at
* **Categorical:** category\_list, country\_code, region, city
* **Outcome:** status (operating/acquired/closed)

Note: Missing values were handled by substituting funding dates logically based on available timestamps.

**3. Preprocessing Steps**

* Dates were parsed into datetime format and missing entries imputed using domain logic (e.g., substituting first funding date when founding date is missing).
* Funding totals were converted to numeric values for analysis.
* Categorical sectors (category\_list) will be enriched for segmentation.

**4. Proposed Analytical Framework**

a. Startup Segmentation

Group startups based on:

* Industry sectors
* Country and region
* Funding amount brackets
* Age of company at first/last funding

This allows us to benchmark risk by peer groups.

b. Risk Factor Analysis

We aim to identify:

* Which sectors have the highest closure rates?
* Is early funding (or lack thereof) linked to increased risk?
* Are there funding patterns (e.g., high # of small rounds) that correlate with failure?

c. Feature Engineering

* Funding velocity = time between first and last funding
* Startup age at funding
* Funding concentration = total amount / # of rounds

These new variables can enrich the credit scoring model.

d. Machine Learning Model

If resources allow, a Random Forest Classifier could be trained on features like:

* Sector, funding size, # rounds, funding timing
* Country/region
* Engineered features from above

The model would classify whether a startup is likely to operate or fail and assign a credit risk score.

**5. Deliverables**

We propose to deliver:

* A dashboard or slide deck showing segment risks and recommended policies
* A simple scoring matrix for credit officers (e.g., low/medium/high risk)
* Optional: ML risk scoring tool

**6. Data Architecture: Bronze–Silver Framework**

To ensure data quality, transparency, and scalability, we adopt a simplified Bronze–Silver Lakehouse architecture inspired by best practices from industry leaders such as element61.

Bronze Layer – Raw Data Ingestion

The Bronze Layer contains the unprocessed data extracted directly from the source. In our case, this includes the original startup\_failures.csv file as downloaded from the public repository. This layer preserves:

* All original columns and raw values
* Incomplete and inconsistent entries
* Full historical fidelity (including anomalies)

The Bronze Layer is not directly used for analysis but serves as a secure, auditable backup and a source for all further transformations.

Silver Layer – Cleaned & Enriched Analytical Data

The Silver Layer contains cleaned, validated, and enriched data that is ready for statistical analysis and modeling. For this project, the Silver Layer includes:

* Standardized datetime variables (founded\_at, first\_funding\_at, last\_funding\_at)
* Currency normalization of the funding\_total\_usd column
* Logical imputation of missing founding and funding dates using a cascading strategy
* Feature engineering:
  + *Startup Age* at time of first and last funding
  + *Funding Velocity*: time between first and last funding rounds
  + *Funding Density*: amount per round
* Categorical enrichment of the category\_list to allow sector-based segmentation
* Outcome Label (status) indicating operational success or failure

This Silver Layer functions as the core analytical dataset for:

* Segmenting startups by risk profile
* Developing a credit risk strategy
* Potentially training machine learning models

Gold Layer – Turning Data into Real-Time Intelligence

The Gold Layer takes our analysis one step further by using real-time and external data to make credit decisions smarter and more up to date. Instead of relying only on historical data, this layer helps us understand what’s happening with startups right now.

For example, we could connect to:

* Funding platforms like Crunchbase or AngelList to see if a startup just raised new money
* News sources or social media to catch mentions of layoffs, scandals, or growth
* Job boards to see if the company is hiring or shrinking
* Patent databases to check for innovation and R&D activity

With this kind of live information, we can update a startup’s risk score every day if needed—seeing which companies are gaining momentum and which might be in trouble.

We also include wider economic and policy trends that affect a startup’s future. If an industry is shrinking (like construction during a downturn) or facing new regulations (like crypto in Europe), the risk of failure increases—even if the startup’s past looks good. On the other hand, startups in sectors like green energy or AgriTech might become safer bets if governments are offering subsidies or support.

In short, the Gold Layer helps us give smarter, fairer, and more future-proof risk scores. Two startups that look similar on paper might actually be very different when you look at what's happening around them. This layer makes sure we don’t miss that.